돼지 사육두수 관측을 위한 표본설계

김 연 중* 유 찬 주**

Key words: 표본설계(sampling design), 절사법(cut-off sampling), Neyman 최적 할당법, 사육두수 추정(prediction of hog population), 양돈농가(hog farming)

ABSTRACT

The purpose of this study is introduce of sampling method and prediction of hog population. Before computing sample size we cut off small farms from the sampling frame. The contribution from this part of the population is at least small in comparison with the remaining population. It may be tempting not to use resources on farms that contribute little to the overall result of the survey. Moreover, this reduces the response burden for these small farms. These farms were excluded from the frame so that the sample size could be reduced without significantly affecting quality.

The sample design is based on 2000 agricultural census data and other related government data. Two stage stratified random sampling is applied for the sample design, where the first stage stratum is a nine geographical area(province), and the second stage is a seven stratum within province. The sample are 388 households for the pig.

1. 서론

연구자가 조사를 통해 모집단의 특성을 찾기 위해서는 전수조사를 하는 것이 가장 바람직하다. 그러나 전수조사가 불가능한 경우에는 표본조사를 실시한다. 표본조사는 전수조사에 비해 조사비용과 조사원이 적게 소요되며, 단기간에 필요한 자료 수집이 가능하고, 모집단의 일부인 표본을 집중

* 한국농촌경제연구원 부연구위원.
** 전북대학교 농업과학기술연구소.

적으로 조사할 수 있어 자료의 질이 전수조사보다 우월할 수도 있다. 이러한 이유로 모집단이 아주 작은 경우가 아니라면 대부분 표본조사를 하고 있다.

이와 함께 연구자가 많이 이용하고 있는 표본조사에 대해 보다 통계적으로 유의한 결과를 얻기 위해서는 연구 대상에 대한 모집단을 규정하고, 모집단을 중심으로 표본추출 틀을 결정한 다음, 표본추출 틀 내에서 자료의 통계적 특성을 고려하여 표본추출 방법을 결정하고, 조사비용과 연구인력 등을 고려하여 표본의 크기를 결정한 후 표본을 추출할 필요가 있다.

특히 표본 설계를 작목에 적용할 경우에 경영 환경의 변화가 심하거나 조사 분석이 어려운 작목에서 활용할 수 있다. 양돈경영의 경우를 살펴보면, 1990년 이후 소규모 사육농가의 감소와 대규모 사육농가의 증가로 인해 급격히 대규모화, 전업화하고 있다. 즉, '90년대에 전체 호수의 0.5%에 불과했던 1,000두 이상 사육농가가 2003년에 와서 19.2%로 증가하였으며, 동기간 전업화율도 38.4%에서 81.8%로 2배 이상 증가하였다. 이와 같이 양돈경영은 타 축종에 비해 급속한 성장을 하고 있지만, 양돈경영의 경우 질병문제로 인해 폐쇄적 경영관리가 이루어지고 있어 조사연구가 어렵다. 따라서 통계적으로 유의성이 있는 표본 설계가 매우 중요하다. 현재 표본 설계와 관련된 조사연구는 관측정보센터에서 이루어지고 있으나 두 가지 측면에서 한계를 가지고 있다. 즉, 표본 농가 선정시 표본 수를 500두 미만, 500-1,000두, 1,000-5,000두, 5,000두 이상 등과 같이 층별 사육농가 수에 비례해서 추출하기 때문에 통계적으로 유의하지 못하다[1]. 그리고 표본 농가로부터 자료를 수집하여 총 돼지 사육두수 추정할 경우 층별 가중치는 사육농가 수를 기준으로 하는 것이 통계적으로 유의하나 층별 두수를 이용하고 있다.

따라서 이 연구에서는 양돈농가를 대상으로 표본조사에 의거하여 연구 결과를 도출하려는 연구자에게 표본 설계 방법과 표본 설계 순서를 소개하고, 선정된 표본으로부터 자료 수집과 분석 방법을 제시하고자 한다. 그리고 표본 설계 자료는 통계적 신뢰성이 높은 농업총조사(2000년) 자료를 이용하고자 한다.

2. 표본 설계의 이론적 검토

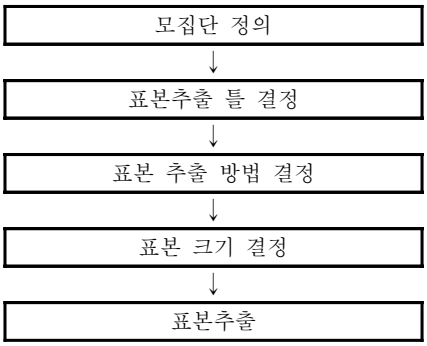
2.1. 표본 설계 순서

표본 설계 순서는 연구 목적에 맞도록 연구자가 모집단을 먼저 규정하고, 모집단을 중심으로 표본추출 틀 결정, 표본추출 틀 내에서 자료의 통계적 특성을 고려하여 표본추출 방법을 결정하고, 신뢰도와 허용오차, 조사비용, 연구인력 등을 고려하여 표본의 크기를 결정한 후 표본을 추출한다 (그림 1 참조).

이 연구에서는 우리나라 돼지 사육두수를 추정하기 위해 표본을 설계하는 것으로

[1] 표본 농가 선정은 농가호당 사육두수별로 층을 설정하고, 층 내에서 분산 정도에 따라 표본의 크기를 조절하는 것이 통계적으로 유의하다.

그림 1. 표본 설계 과정



2000년도 농업총조사 자료를 돼지의 모집단으로 정의하고자 한다. 조사 모집단이 확정되면 연구 대상이나 표본단위가 수록된 표본추출 틀(sampling frame)을 설정해야 한다.

표본추출 틀은 모집단의 구성요소[2]를 모두 포함하되 어떠한 요소도 이중으로 포함하지 않을 때 가장 적합한 추출 틀이라 한다. 표본추출 틀의 오차는 모집단과 표본추출 틀이 완벽하게 일치하지 않을 때 발생하는 오차이며, 이를 줄이기 위해서는 표본추출 틀을 모집단에 맞게 재구성해야 하며, 자료 수집 과정에서 부적절한 문항을 제거하거나 자료 분석 과정에서 수집된 자료의 중요도에 따라 가중치를 적용시켜 조정한다.

표본추출 방법은 표본조사를 실시하기 위해서 가장 중요한 과정으로 조사 대상 즉 모집단에서 표본을 어떻게 추출할 것인가이다. 가장 먼저 결정해야 하는 것은 확률표본으로 추출하느냐, 비확률표본으로 추출하느냐이다.

[2] 모집단의 성질인 양돈농가의 호당 사육두수, 모돈수, 자돈수, 축사의 크기, 지역 등을 많이 포함할수록 좋은 표본추출 틀을 구성할 수 있다.

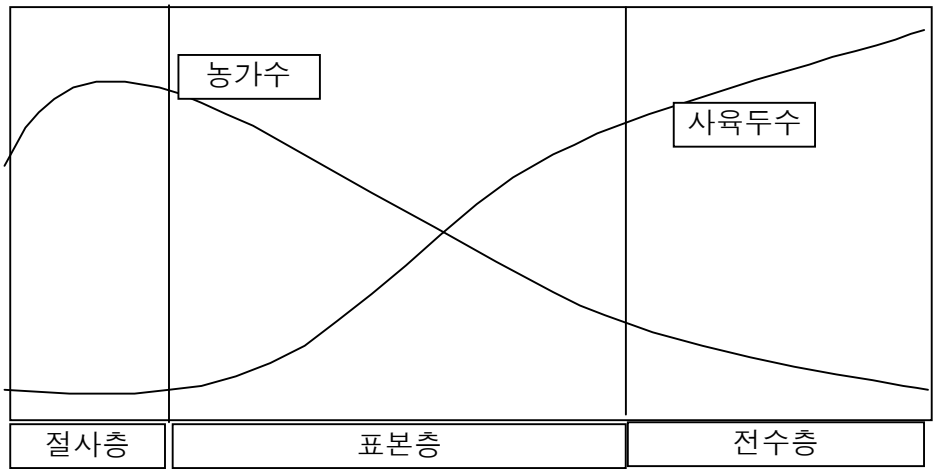
표본의 크기는 표본조사에서 먼저 제기되는 과제 중의 하나이다. 즉, 표본을 어느 정도로 구성해야 타당하고 적절할 것인가 하는 문제는 조사자 입장에서 가장 큰 관심사이다. 표본의 크기는 조사비용과 시간에 밀접한 관계가 있고, 통계량의 수준(신뢰수준, 허용오차범위), 모집단의 동질성 정도 등과 관련이 깊다.

2.2. 표본추출 방법

표본 추출 방법을 크게 두 가지로 나누면 확률표본추출과 비확률표본추출이다. 확률표본추출은 무작위 추출을 전제로 하며, 확률이론에 입각하여 통계적 추론이 가능한 것으로 모집단을 구성하고 있는 개별요소가 표본에 포함될 확률이 동일한 표본추출 방법이다. 대표적인 방법은 단순임의추출, 층화추출, 집락추출, 계통추출 등이 있다.

비확률추출법은 모집단의 구성 요소가 표본으로 추출될 확률이 동일하지 않은 표본추출 방법이다. 대표적인 비확률표본추출법은 편의표본추출, 할당표본추출, 유의표본추출, 눈덩이표본추출 등이 있다. 비표본확률추출의 단점은 표본으로 선택될 확률이 알려져 있지 않기 때문에 선정된 표본이 모집단을 대표한다고 할 수 없다. 그러나 비확률표본추출은 적용하기가 간단하고 시간과 비용이 적게 들며, 통계적으로 복잡하지 않기 때문에 사용자가 많다. 그러나 이 연구에서는 가장 많이 사용하고 있는 확률표본추출법 내에서 층화추출방법을 이용하였다.

그림 2. 절사층을 포함한 모집단의 표현



2.2.1. 층화추출

모집단 내의 상이하고 이질적인 원소들이 중복되지 않도록 동질적이고 유사한 원소들로 묶은 여러 개의 부모집단으로 나누어 층을 형성한다. 층화추출은 층내 개체들은 통계치(평균, 분산)가 서로 유사하나 층간은 서로 다르다는 것을 가정하고 있다.

각 층내에서 단순임의추출(simple random sampling)하면 층화임의추출(stratified random sampling)이고, 층화변수가 하나이면 일원층화추출(one way stratified sampling)이고, 층화변수가 두 개이면 이원층화추출(two way stratified sampling)이다. 만약 사육두수를 층화변수로 이용하여 200두 미만, 200~500두, 500~1,000두, 1,000~3,000두, 3,000두 이상으로 구분하여 표본을 추출하면 일원층화추출이다. 이를 지역별로 층화하면 이원층화, 축산농가의 노동력수를 동시에 층화했을 경우 삼원층화추출이다.

또한 층화변수가 대분류인 경우, 대분류를 중분류로 나누어 층화하면 2단 층화추출(two stage stratified sampling)이고, 중분류를 다시 소분류로 층화하면 3단 층화추출(three stage stratified sampling)이라고 한다.

2.2.2. 층별할당(Neyman의 최적할당, 절사법)

층별로 표본의 수를 할당하는 방법으로 층내에서 같은 수의 표본을 할당하는 방법을 균등할당이라 하고, 층의 크기에 비례하여 할당하는 방법을 비례할당, 층의 크기와 분산 크기의 곱에 비례하여 할당하는 방법을 최적할당이라 한다.

층화추출 시 대부분의 경우 모집단을 두 개의 층으로 나누고, 전수조사가 시행되는 층은 전수층, 표본조사가 이루어지는 층은 표본층이라 하며, 왜도가 심한 경우 절사층, 전수층 그리고 하나 이상의 표본층으로 구성되는 경우가 있다(그림 2 참조). 이와 같이 여러 개 층 가운데 절사층이 포함되어 있으면 절사법(cut-off sampling)을 이용하여 최적할당을 한다.

2.3. 표본 크기 결정

표본 크기를 몇 개로 할 것인가 하는 문제는 통계분석에서 중요한 문제의 하나이다. 표본 설계 이후 표본조사를 실시하여 분석 결과를 얻었을 때, 신뢰수준과 표본오차를 어느 정도로 할 것이냐에 따라 표본수가 정해진다. 또한 표본조사 비용과 시간이 어느 정도 있느냐에 따라 다를 수 있다.

2.3.1. 왜도가 심한 경우

양돈농가의 표본의 크기를 결정하기 위해 2000년도 시행된 농업총조사 자료를 이용하였다. 자료의 분석 결과에서 사육규모에 대한 왜도(skewness)[3]가 19이상으로 크고, 지역별 특성치에 대한 분산이 서로 다르게 나타나 Neyman의 최적할당식을 적용하였다. 표본의 수는 신뢰수준(95% 또는 90%)과 허용오차(±0.05~±0.06)에 따라 다르다. 표본의 크기 결정은 (식 1)에 의해 정한다.

$$n = \frac{\left(\sum_{h=1}^{L} N_h S_h\right)^2}{N^2 D + \sum_{h=1}^{L} N_h S_h^2} \qquad \text{(식 1)}$$

여기에서 n은 총 표본의 크기, n_h는 h층의 표본 크기, N_h는 h층의 부모집단크기, W_h는 h층의 가중치, S_h는 h층의 모표준편차, N은 모집단의 크기, D는 (허용오차/신뢰수준)[2]이다. 여기서 허용오차와 신뢰수준에 따라 표본 수가 많아지기도 하고 적어지기도 한다.

2.3.2. 절사법

모집단의 분포가 심한 왜도를 보이거나 소수의 모집단 요소들이 모집단 총계의 대부분을 차지하는 경우 Neyman의 최적할당법은 지나치게 큰 표본 크기를 할당하여 많은 조사비용이 요구될 뿐만 아니라 비표본오차(non-sampling error)가 크게 되어 추정치의 신뢰를 감소시킬 수 있다.

이와 같은 현상은 농업 생산량 조사, 사업체조사(business survey)에서 심하게 나타난다. 이러한 모집단에 대한 표본 크기의 할당은 모집단을 전수층(take-all stratum)과 표본층(take-some stratum)으로 구분하여 주어진 정도와 신뢰계수를 만족하는 표본의 크기를 결정하는 방법인 절사법[4]을 응용함으로써 Neyman의 최적할당보다 적은 표본 크기를 결정할 수 있게 된다.

2.3.3. 상대오차 개념

모집단 추정치 혹은 모평균 추정치의 상대오차를 통제함으로써 표본의 크기를 결정할 수 있다. 단순임의추출(simple random sampling)에 의해서 모평균 추정을 할 경우 표본의 크기는 다음과 같은 (식 2)에 의한다. 단, 단순임의 추출하에서 $\bar{y}$가 정규분포를 따른다고 가정한다.

---

[3] 왜도는 평균에 대한 비대칭 정도를 나타내는 것으로 왜도가 0.5보다 크면 분포의 꼬리가 양의 값 쪽으로 치우쳐 있고, -0.5보다 작으면 이와 반대 방향으로 치우친 것이다. 양돈농가의 사육규모에 대한 왜도는 19.1로 매우 큰 것으로 나타났다.

[4] Deming(1960)에 의해 처음으로 제안되어 이후 많은 연구에 이용되고 있다.

$$r\overline{Y} = t\sigma_{\bar{y}} \quad \text{(식 2)}$$

$$r = \frac{t\sigma_{\bar{y}}}{\overline{Y}}$$

$$\sigma_{\bar{y}} = \sqrt{\frac{N-n}{N}}\frac{S}{\sqrt{n}}$$

$$r = \frac{t\sqrt{\frac{N-n}{N}}\frac{S}{\sqrt{n}}}{\overline{Y}}$$

여기서 γ: 상대오차, $\overline{Y}$: 모평균, $\overline{y}$: 표본평균, $\sigma_{\bar{y}}$: 모분산, n: 표본 크기, N: 모집단 크기, t : 분포값 이다.

2.4. 표본의 최적할당

신뢰수준과 표본오차에 따라 총표본수를 구한 후에 층별로 표본 수를 할당해야 한다. 이를 위해서는 Neyman의 최적할당 방법을 이용하는데, 할당방법은 (식 3)과 같다.

$$n_h = \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} \times n \quad \text{(식 3)}$$

$$= \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \times n$$

여기서 n_h: h 번째 층의 표본 크기, W_h: h 층의 가중치, h 번째 부모집단의 크기, N_h: h 번째 부모집단의 크기, S_h: h 번째의 부모집단의 표준편차이다. 지금까지는 일반적으로 층별 총 사육두수 또는 층별로 총 농가 수에 비례해서 표본 수를 할당하였다. 그러나 통계적으로 유의성을 갖기 위해서 표본 수는 층별로 사육농가 수와 사육두수의 표준편차를 곱해서 나온 값을 가중치로 이용하여 할당해야 한다.

3. 양돈농가 표본설계

3.1. 모집단 정의

표본 설계를 위한 모집단은 농업총조사(2000년) 자료의 양돈농가 전체를 기본으로 하였다. 필요한 자료 중 농업총조사에서 조사되지 않은 정보는 보조자료를 이용하였다. 보조자료는 국립농산물품질관리원의 “가축통계”를 이용하였다.

농업총조사에 나타난 자료를 정리하면 돼지 총 사육두수는 7,059,309두, 사육 농가수는 24,126호, 호당 평균 사육두수는 292두이다. 사분위수를 이용하여 자료를 정리해 보면 제1사분위수에 해당하는 사육두수가 1~2마리에 불과하였다. 즉 사육두수를 1두부터 최대사육두수인 51,000두까지 크기 순으로 정렬했을 때, 하위 25%에 해당하는 농가가 2두 이하를 사육한다는 의미이다. 이는 24,126 농가의 25%에 해당하는 6,054 농가가 2두 이하를 사육한다는 것이며, 6,054농가가 사육하는 총 돼지 사육두수는 최대 12,000마리 이하이며, 이는 총 사육두수 7,059,309의 0.17%에 해당된다. 제2사분위수는 3~15두, 제3사분위수는 16~300두, 300두 이상이 제4사분위수에 해당되는 것으로 나타났다. 이 결과에 따르면 양돈농가의 호당규모에 대한 왜도가 19.1로 나타나 한쪽으로 심하게 치우쳐 있다는 것을 알 수 있다.

3.2. 표본추출 틀

모집단이 확정되었기 때문에 표본추출

틀(sampling frame)을 설정해야 한다. 표본추출 틀로부터 최종적인 표본이 추출된다. 좋은 표본추출 틀은 모집단과 일치하는 것이 좋으나, 대부분 모집단과 표본추출 틀이 정확하게 일치하지는 않는다.

양돈농가와 같이 왜도가 심한 경우 표본추출 틀을 확정하는 데 통계적으로 유의한 범위 내에서 절사하는 방법이 있다. 절사방법으로 이 연구에서는 일정 규모 이하의 사육농가를 절사하고, 사육두수 추정에 영향을 적게 주는 비주산지 지역을 표본추출 틀에서 절사하였다.

3.2.1. 규모에 대한 절사

농업총조사 자료에 따르면 소규모 양돈농가가 많고, 이는 총 사육두수에 차지하는 비중이 아주 적다. 표본 설계의 목적이 돼지 사육두수의 총계 추정이므로 소규모 사육농가를 조사 모집단에서 절사하더라도 사육두수 추정에 큰 문제가 없다.

소규모 양돈농가의 절사방법은 모집단 내의 사육농가를 사육두수 크기 순으로 정렬한 후 200두 미만의 농가를 절사하였다. 이는 총 사육두수에서 차지하는 비중이 통계적으로 유의수준 범위 내에 있기 때문에 이를 기준으로 절사하였다. 즉 200두 미만을 절사할 경우 조사모집단의 크기는 목표모집단의 99.3%에 해당하는 반면에 조사모집단의 농가 수는 무려 51.2%가 감소하는 효과가 있기 때문이다.

실제로 양돈농가의 절사 현황을 <표 1>에서 보면, 200두 이하 사육농가를 절사한 경우, 모집단 24,216농가에서 16,870농가가 절사되어 조사모집단은 7,256농가로 전체 농가에서 약 70%가 줄었다. 그러나 총 사육두수 7,059,309에서 348,580두가 절사되어 조사 사육두수는 6,710,729이며 전체의 95.1%이다.

결과적으로 200두 미만의 농가를 절사했을 때 조사 모집단수는 큰 폭으로 줄었으나, 조사 사육두수는 4.9% 밖에 줄지 않았다. 이는 모집단의 총계를 통계적으로 추정하면서 표본의 크기를 줄일 수 있을 뿐만 아니라 소규모 사육농가들의 신규 혹은 폐업(birth & death)으로 인한 무응답의 발생을 사전에 배제하여 비표본오차도 줄이는 효과가 있다.

3.2.2. 비주산지 절사

양돈농가의 비주산지는 조사 대상 지역에서 절사하였다. 이는 조사비용과 추정의 효율을 고려할 때 필요한 조치이다.

비주산지의 사육두수는 총계에서 차지하는 비중이 적어 이를 조사 모집단에서 제외한다는 원칙과 부합하는 것이다. 이들 지역은 서울, 인천, 대전, 대구, 광주, 부산, 울

표 1. 양돈농가의 절사 현황

절사 두수	사육농가				사육두수			
	전체	절사	조 사 모집단	모집단 비 율	전체	절사	조 사 모집단	모집단 비 율
200	24,126	16,870	7,256	30.1	7,059,309	348,580	6,710,729	95.1

산 등이며, 조사표본은 서울·인천의 경우 경기에, 대전은 충남에, 대구는 경북에 광주는 전남에, 부산·울산은 경남에 할당하였다.

3.3. 층 결정

표본 설계의 목적이 사육두수의 총계 추정이므로 사육두수를 층화변수로 이용하였다. 층의 경계는 <표 2 >에서 $CUM\ \sqrt{f(y)}$ 방법을 이용하였다.

또한 양돈농가 간에 동일한 두수를 사육하는 경우가 있으므로 규모가 비슷한 농가를 먼저 계급화하고, 낮은 계급에서 높은 계급으로 정렬하면 된다. 계급의 빈도는 루트로 계산 후 누적 총계를 구할 수 있다. 누적 총계를 기준으로 2개 층, 3개 층, 4개 층 그리고 여러 개 층으로 나눌 수 있는데, 층의 수가 많으면 많을수록 분산은 작아지지만 표본선정이나 분석하기가 어렵다. 따라서 이 연구에서는 분산이 적은 7개 층으로 구분했으며, 층의 경계를 구하는 것은 (식 4)와 같다.

$$L_1 = \frac{\sum_{i=1}^{k}\sqrt{f_i}}{L},\quad L_2 = \frac{2\sum_{i=1}^{k}\sqrt{f_i}}{L},\ \text{(식 4)}$$

$$\cdots\ L_{(L-1)} = \frac{(L-1)\sum_{i=1}^{k}\sqrt{f_i}}{L}$$

여기에서 L은 층의 수이고 L_1은 첫 번째 층과 두 번째 층의 경계점이다. <표 3>을 보면 200~209마리까지 사육하는 농가가 682호이고, 이들이 136,494두를 사육하고 있다. 26은 농가 수 682호를 루트로 계산한 것이다. 이 표에서 총 누적 루트 는 709.04이며, (식 4)를 이용하여 다음과 같이 층의 경계를 구하였다.

$$L_1 = \frac{\sum_{i=1}^{k}\sqrt{f_i}}{7} = \frac{709.3043747}{7} = 101.3292$$

$$L_2 = \frac{2\sum_{i=1}^{k}\sqrt{f_i}}{7} = \frac{2\times 709.3043747}{7} = 202.6584$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$L_6 = \frac{6\sum_{i=1}^{k}\sqrt{f_i}}{7} = \frac{6\times 709.3043747}{7} = 607.9752$$

표 2. 층 구분 방법

계급	$f(y)$ (빈도)	$\sqrt{f(y)}$ ($\sqrt{}$ 빈도)	$CUM\ \sqrt{f(y)}$ (누적 $\sqrt{}$ 빈도)
1	f_1	$\sqrt{f_1}$	$\sqrt{f_1}$
2	f_2	$\sqrt{f_2}$	$\sqrt{f_1}+\sqrt{f_2}$
3	f_3	$\sqrt{f_3}$	$\sqrt{f_1}+\sqrt{f_2}+\sqrt{f_3}$
4	f_4	$\sqrt{f_4}$	$\sqrt{f_1}+\sqrt{f_2}+\sqrt{f_3}+\sqrt{f_4}$
5	f_5	$\sqrt{f_5}$	$\sqrt{f_1}+\sqrt{f_2}+\sqrt{f_3}+\sqrt{f_4}+\sqrt{f_5}$
.	.	.	.
.	.	.	.
k	f_k	$\sqrt{f_k}$	$\sqrt{f_1}+\sqrt{f_2}+\sqrt{f_3}+\sqrt{f_4}+\sqrt{f_5}+\ldots+\sqrt{f_k}$
		$\sum_{i=1}^{k}\sqrt{f_i}$	

층 1의 경계점을 <표 3>에서 보면 101.3292로 '누적'의 98.52228과 104.2668 사이에 있음을 알 수 있다. 따라서 정확한 층 1과 층 2의 경계점은 300두와 310두 사이에 존재한다고 볼 수 있으나, 농가 수가 310두 쪽에 많아 310두를 경계선으로 정하였다.

층 2와 층 3의 경계점은 202.6584로 누적의 185.9819와 213.4045에 존재하나 편의상 500두로 정하였다. 층 3과 층 4의 경계점은 303.9876으로 누적의 288.5753가 307.8626에 존재하나 편의상 700두로 정하였다. 층 4와 층 5의 경계점도 같은 방식으로 정하였으며, 층 6과 층 7의 경계는 607.9752로

표 3. 양돈농가 층 경계

계급(두수)	농가수	사육두수	농가 수 root	누적 root
200-209	682	136,494	26.11512971	26.11513
210-219	29	6,333	5.385164807	31.50029
220	44	10,097	6.633249581	38.13354
300	14	4,297	3.741657387	98.52228
310	33	10,533	5.744562647	104.2668
320	5	1,642	2.236067977	106.5029
⋮	⋮	⋮	⋮	⋮
480	6	2,933	2.449489743	185.9819
490	752	376,000	27.4226184	213.4045
500	9	4,572	3	216.4045
⋮	⋮	⋮	⋮	⋮
680	6	4,133	2.449489743	288.5753
690	372	260,399	19.28730152	307.8626
700	2	1,414	1.414213562	309.2768
⋮	⋮	⋮	⋮	⋮
970	2	1,975	1.414213562	397.3036
980	680	680,000	26.07680962	423.3804
990	3	3,008	1.732050808	425.1124
⋮	⋮	⋮	⋮	⋮
1,440	1	1,440	1	502.7302
1,450	2	2,900	1.414213562	504.1445
1,500	325	487,500	18.02775638	522.1722
⋮	⋮	⋮	⋮	⋮
2,400	9	21,600	3	602.7912
2,500	93	232,500	9.643650761	612.4348
2,534	2	5,067	1.414213562	613.849
⋮	⋮	⋮	⋮	⋮
30,000	3	90,000	1.732050808	708.3044
50,000	1	50,000	1	709.3044
	7,256	6,710,729	709.3043747	

표 4. 층별 지역별 농가 수 및 사육두수

사육규모		1층 200~310	2층 311~500	3층 501~700	4층 701~990	5층 991~1,450	6층 1,451~2,500	7층 2,500~	전 체 농가수	평 균 사육두수
농가수	경기	335	399	244	179	260	213	72	1,702	943
	강원	74	53	31	42	43	26	10	279	898
	충북	76	73	56	37	58	42	19	361	1,053
	충남	299	293	162	123	219	179	42	1,316	907
	전북	225	174	101	82	152	82	24	840	828
	전남	189	139	109	77	115	77	24	730	850
	경북	220	171	136	108	138	112	44	930	969
	경남	204	179	100	87	151	92	30	843	877
	제주	36	27	25	15	63	73	15	255	1,273
계		1,658	1,508	964	750	1,200	896	280	7,256	925

누적의 602.7912와 612.4348에 존재하여 2,500두 이상으로 정하였다.

조사 모집단 7,256농가의 층별 분포를 나타낸 것이 <표 4>이다. 1층은 1,658농가로 가장 많고, 경기 지역의 조사 모집단이 많은 것으로 나타났다. 호당 평균 사육두수는 925두인데, 제주지역은 1,273두, 충북 1,053두, 경북 969두의 순이며, 전북지역의 호당 사육두수가 828두로 가장 적은 것으로 나타났다.

3.4. 표본 수 결정과 할당

양돈농가의 층의 경계가 결정되었고, 표본의 수는 신뢰수준(95% 또는 90%)과 허용오차(±0.05~±0.06)에 따라 다르다. 표본의 크기는 Neyman 할당식인 (식 1)에 의해 정하였다.

전북지역의 표본 크기를 설정하는 경우를 살펴보면 (식 1)에 전북 지역 양돈농가의 현황을 대입하고, 신뢰수준 95%에 허용오차 5%를 적용하여 전북지역의 표본의 크기를 구하면 39개가 된다.

$$n = \frac{134,857^2}{696,076^2 \times 0.000650771 + 157,164,377}$$

$$= 39\text{개 표본수}$$

표 5. 양돈 표본 농가의 최적할당(전북)

층	총 사육두수	표준편차(S_h)	농가수(N_h)	N_hS_h	$N_hS_h^2$	할당농가
1	56,721	45	225	10,125	455,625	3
2	76,274	57	174	9,918	565,326	3
3	62,584	60	101	6,060	363,600	2
4	68,423	62	82	5,084	315,208	1
5	166,863	126	152	19,152	2,413,152	6
6	149,801	311	82	25,502	7,931,122	7
7	115,410	2,459	24	59,016	145,120,344	17
계	696,076	-	840	134,857	157,164,377	39

전북지역의 총표본수가 39개로 결정되었으므로 층별로 표본 수를 할당해야 한다. 관심변수가 사육두수이므로 층별 크기가 다를 뿐만 아니라 층별 분산이 서로 다르다. Neyman의 최적할당방법은 표본변동계수를 줄이는 가장 효율적인 방법으로 Neyman의 층별 최적할당은 (식 3)에 의해 다음과 같이 구하였다.

$$n_1 = 39 \times \frac{10,125}{134,857} = 3$$

$$n_2 = 39 \times \frac{9,918}{134,857} = 3$$

$$n_3 = 39 \times \frac{6,060}{134,857} = 2$$

$$n_4 = 39 \times \frac{5,084}{134,857} = 1$$

$$n_5 = 39 \times \frac{19,152}{134,857} = 6$$

$$n_6 = 39 \times \frac{25,502}{134,857} = 7$$

$$n_7 = 39 \times \frac{59,016}{134,857} = 17$$

여기서 낮은 층의 농가 수가 많음에도 불구하고 표본 수가 대체로 적고, 7층의 경우 농가 수는 적은데 표본의 수가 많다. 7층은 1층에 비해 호당 사육두수 표준편차가 크기 때문에 표본 수가 많은 것이다. 지금까지 표본의 크기를 결정할 때 농가 수에 비례해서 정하는 것이 일반적이었으나 통계적으로 유의하기 위해서는 표준편차의 크기에 따라 표본 수를 정하는 것이 바람직하다. 지역별 층별 표본의 크기는 <부록>에 정리하였다.

3.5. 표본 설계 결과

우리나라 총 돼지 사육두수를 추정하기 위해 신뢰수준별, 지역별 표본 농가수를 구한 것이 <표 6>이다. 신뢰수준 90%, 95%와 허용오차 ±5%, ±6%에 따라 표본 수는 각각 332개, 277개, 388개, 331개이다. 신뢰수준 95%에 허용오차 ±5를 기준으로 할 때 필요 표본 수는 388농가로 지역별로 보면 경기·인천이 72, 강원 19, 충북 27, 충남·대전 60, 전북 39, 전남·광주 40, 경북·대구·울산 59, 경남·부산 45, 제주 27 농가이다.

신뢰수준 95%와 허용오차 5% 범위 내에서 총표본수와 지역별·층별 표본 수를 구한 것이 <표 7>이다. 전체 24,216농가중에서 388농가의 표본이 통계적으로 유의하다고 볼 수 있다. 조사비용과 시간에 따라 표본의 수를 늘리거나 줄일 수 있다.

표 6. 지역별 표본 농가수

신뢰수준	허용오차	경기·인천	강원	충북	충남·대전	전북	전남·광주	경북·대구·울산	경남·부산	제주	계
90%	±5%	57	18	26	53	30	33	53	37	25	332
	±6%	44	17	24	46	23	25	46	29	23	277
95%	±5%	72	19	27	60	39	40	59	45	27	388
	±6%	57	18	26	53	30	32	53	36	26	331

표 7. 표본 농가의 지역별 층별 할당

층	경기·인천	강원	충북	충남·대전	전북	전남·광주	경북·대구·울산	경남·부산	제주	계
1	2	1	1	2	3	2	2	2	1	16
2	4	1	1	2	3	2	2	3	1	17
3	2	0	1	1	2	2	1	1	1	11
4	2	1	0	1	1	1	1	1	1	9
5	5	1	1	4	6	4	3	5	2	32
6	12	3	2	8	7	8	7	8	6	59
7	46	12	22	42	17	20	44	24	15	245
계	72	19	27	60	39	40	59	45	27	388

4. 사육두수 추정

우리나라 총 돼지 사육두수 추정을 위해서는 지역의 사육두수를 추정한 후 이를 합계하여 구할 수 있다. 총계 추정은 (식 5, 6)을 이용하였다.

$$\hat{\tau}_{\text{전국}} = \sum_{t}^{n} \hat{\tau}_{\text{지역}} \quad \text{(식 5)}$$

$$\hat{\tau}_{\text{전북}} = N_{\text{전북}} \bar{y}_{\text{전북}}$$

$$\bar{y}_{\text{전북}} = \sum_{h=1}^{7} W_{\text{전북}h} \bar{y}_{\text{전북}h} \quad \text{(식 6)}$$

$$= \sum_{h=1}^{7} \frac{N_{\text{전북}h}}{N_{\text{전북}}} \bar{y}_{\text{전북}h}$$

여기서

$\hat{\tau}_{\text{전국}}$: 전국 총 돼지사육두수 추정치

$\hat{\tau}_{\text{전북}}$: 전북 지역의 총계추정치

$N_{\text{전북}h}$: 전북 지역에서의 h번 층의 부모집단 크기

$W_{\text{전북}h}$: 전북 지역의 층별 가중치

$\bar{y}_{\text{전북}h}$: 전북 지역에서의 h번 층의 표본평균

h: 층을 나타내는 첨자

지역의 사육두수를 추정하기 위해 전북지역의 사육두수를 예로 추정해 보면 다음과 같다. 전북지역의 총 양돈농가 중에서 200두 미만을 제외한 부모집단 수는 840농가이다. 이들이 사육하고 있는 총 사육두수는 696,076이고 호당 평균 사육두수는 828이다.

전북지역의 표본을 층별로 보면 1~2 층은 각 3농가, 3층은 2농가, 4층은 1농가, 5층 6농가, 6층 7농가 그리고 7층은 17농가로 총 39개의 표본 농가이다. 표본 농가로부터 사육두수를 조사한 후 전북지역의 총 사육두수가 몇 두 인지 (식 6)을 이용하여 추정할 수 있고, 실제 자료를 대입하여 계

표 8. 전북 돼지 사육두수 현황

층	총 사육두수	농가수(Nh)	호당평균 사육두수	할당농가
1	56,721	225	252	3
2	76,274	174	438	3
3	62,584	101	620	2
4	68,423	82	834	1
5	166,863	152	1,096	6
6	149,801	82	1,825	7
7	115,410	24	4,809	17
계	580,666	816	828	39

산해 보았다.

평균 사육두수 ($\bar{y}_{\text{전북}}$)

$$= \frac{225}{840} \times 252 + \frac{174}{840} \times 438 + \cdots + \frac{24}{840} \times 4,805$$

= 828.361, 총 사육두수($\hat{\tau}_{\text{전북}}$)

= 840 × 828.361 = 696,076

전북지역의 표본조사 결과 1층의 사육농가의 평균두수가 280두, 2층 450두, 3층 620두, 4층 850두, 5층 1,100두, 6층 1,850두, 7층 4,900두이면 (식 6)을 적용하여 계산하면 총 사육두수는 2.1% 증가한 710,748두로 추정된다. 전국 사육두수 추정은 전북지역 사육두수 추정과 같이 지역별로 계산한 후 더하면 된다. 우리나라 돼지 사육두수 총계 추정치는 신뢰수준 95%에 허용오차 5% 유의성을 가지고 있다고 말할 수 있다.

5. 요약 및 결론

조사를 통해 연구자의 논리를 전개하기 위해서는 표본 설계, 표본조사 그리고 표본조사 결과의 추정과정 등 단계적 접근이 필요하다. 표본 설계 순서는 모집단을 먼저 규정하고, 표본추출 틀 결정, 표본추출 틀에서 자료의 특성을 충분히 고려하여 표본추출 방법을 결정하고, 신뢰수준과 허용오차의 크기, 조사비용, 조사기간, 연구인력 등을 고려하여 표본 크기를 결정한 후 표본을 추출하는 것이 통계적으로 유의한 결과를 도출할 수 있다.

이 연구에서는 양돈농가를 대상으로 2000년 농업총조사 자료를 이용하여 표본설계 방법을 제시하였다. 모집단을 분석한 결과, 호당 평균 돼지 사육두수의 왜도가 심해, 총계 추정에 영향을 미치지 않은 200두 미만의 사육농가는 절사하였고, 200두 이상 농가를 표본추출 틀로 결정하였다.

표본추출 틀로부터 규모가 비슷한 농가를 먼저 계급화하고, 낮은 계급에서 높은 계급으로 정렬한 후 층을 구분하였다. 층은 많을수록 분산이 적지만 표본선정과 조사분석 시에 많은 노력이 필요하므로 7개 층으로 구분하였다. 또한 지역의 특성을 반영한 표본을 선정하기 위해 비주산지는 절사하였고, 주산지에서 표본을 추출하였다.

표본 수는 조사비용과 조사기간에 영향을 미치며, 표본조사 결과의 신뢰도와 밀접한 관계가 있기 때문에 지역별로 Neyman의 할당 식을 이용하였고, 층별로 다시 할당하여 표본을 추출하였다. 그 결과 전체 24,216농가 중에서 388농가를 표본으로 선정했으며, 통계적으로는 신뢰도 95%, 표본오차 5%이다.

끝으로 표본으로 조사된 자료를 이용하여 우리나라 총 돼지 사육두수를 추정할 수 있다. 즉 지역별(도별)로 사육두수를 추정한 후 지역의 사육두수 추정치를 더하면 우리나라 전체 사육두수를 추정할 수 있다.

부 록

부록. 지역별 층별 표본 수 할당(신뢰수준 95%, 허용오차 5%)

지역	층	총두수	표준편차(S_h)	농가수(N_h)	N_hS_h	가중치	표본수
경기·인천	1	83,216	46	335	15,410	0.0339888	2
	2	180,601	56	399	22,344	0.0492826	4
	3	157,864	54	244	13,176	0.0290614	2
	4	148,690	55	179	9,845	0.0217144	2
	5	285,413	120	260	31,200	0.0688157	5
	6	388,043	346	213	73,698	0.1625506	12
	7	361,830	3,996	72	287,712	0.6345865	46
							72
강원	1	18,287	45	74	3,330	0.0451403	1
	2	23,822	64	53	3,392	0.0459808	1
	3	19,850	56	31	1,736	0.0235326	0
	4	34,756	62	42	2,604	0.0352989	1
	5	47,000	130	43	5,590	0.0757761	1
	6	47,150	333	26	8,658	0.1173648	2
	7	59,500	4,846	10	48,460	0.6569066	12
							19
충북	1	18,966	46	76	3,496	0.0213529	1
	2	32,853	56	73	4,088	0.0249687	1
	3	35,513	57	56	3,192	0.0194961	1
	4	31,176	54	37	1,998	0.0122034	0
	5	63,562	112	58	6,496	0.0396763	1
	6	76,500	313	42	13,146	0.0802932	2
	7	121,400	6,911	19	131,309	0.8020095	22
		258,570			32,416		27
충남·대전	1	74,289	46	299	13,784	0.0313536	2
	2	133,195	54	293	15,767	0.0360678	2
	3	103,598	58	162	9,347	0.0214191	1
	4	101,134	50	123	6,199	0.0140195	1
	5	239,934	126	219	27,485	0.0629032	4
	6	325,850	318	179	56,994	0.1297592	8
	7	216,347	7,358	42	309,049	0.7044776	42
							60

지역	층	총두수	표준편차(S_h)	농가수(N_h)	N_hS_h	가중치	표본수
전북	1	56,721	45	225	10,125	0.0750795	3
	2	76,274	57	174	9,918	0.0735446	3
	3	62,584	60	101	6,060	0.0449365	2
	4	68,423	62	82	5,084	0.0376992	1
	5	166,863	126	152	19,152	0.1420171	6
	6	149,801	311	82	25,502	0.1891040	7
	7	115,410	2,459	24	59,016	0.4376191	17
							39
전남·광주	1	46,997	44	189	8,316	0.0616146	2
	2	61,738	58	139	8,062	0.0597327	2
	3	68,017	61	109	6,649	0.0492635	2
	4	63,890	56	77	4,312	0.0319483	1
	5	125,650	125	115	14,375	0.1065067	4
	6	140,755	334	77	25,718	0.1905489	8
	7	113,000	2,814	24	67,536	0.5003853	20
							40
경북·대구·울산	1	54,641	46	220	10,120	0.0294186	2
	2	76,604	56	171	9,576	0.0278372	2
	3	85,420	58	136	7,888	0.0229302	1
	4	89,957	58	108	6,264	0.0182093	1
	5	151,830	126	138	17,388	0.0505465	3
	6	205,000	346	112	38,752	0.1126512	7
	7	237,350	5,773	44	254,012	0.7384070	44
							59
경남·부산	1	49,760	46	204	9,384	0.0545769	2
	2	80,200	57	179	10,203	0.0593401	3
	3	63,770	55	100	5,500	0.0319877	1
	4	71,807	57	87	4,959	0.0288413	1
	5	65,753	133	151	20,083	0.1168017	5
	6	168,170	321	92	29,532	0.1717566	8
	7	139,770	3,076	30	92,280	0.5366957	24
							45
제주	1	9,060	47	36	1,692	0.0163232	1
	2	12,570	52	27	1,404	0.0135448	1
	3	16,725	46	25	1,150	0.0110944	1
	4	12,600	54	15	810	0.0078143	1
	5	69,500	136	63	8,568	0.0826580	2
	6	133,730	299	73	21,827	0.2105715	6
	7	58,050	4,547	15	68,205	0.6579937	15
							27

참 고 문 헌

김경필 외 6인. 2004.『소비자패널 표본 설계 및 구축』. P074. 한국농촌경제연구원.

김연중 외 5인. 2004.『농업관측 품목별 표본 농가 재설계 연구』. M055. 한국농촌경제연구원.

농림부. 2001.『주요작물 지역별 재배동향』. 국립농산물품질관리원.

농림부. 2002.『농업총조사』. 국립농산물품질관리원.

농림부. 2002.『작물통계』. 국립농산물품질관리원.

농림부. 2003.『가축통계』. 국립농산물품질관리원.

한근식 외 9인. 2002.『조사 방법의 이해』. 교우사.

한근식. 1999.『조사연구방법론』. 경문사.

Cochran. W.G. 1977. *Sampling Technique*. 2nd. ed. Wiley & Sons.

Deming. 1960. *Sampling design in business research*. Wiley & Sons.

Hidiroglou. M.A.. 1986. "The construction of a Self Representing Stratum of Large Units in Survey Design." *The American Statistician* 40.

Little. 1982. *Statistical Analysis with Missing Data*. Wiley & Sons: New York.

Mining. 2001. "Quarrying and Manufacturing Proceedings of Statistics." Canada Symposium.

Rubin. 1987. *Multiple Imputation for Nonresponse in Survey*. Wiley & Sons: New York.

Sharon L. Lohr. 1999. *Sampling: Design and Analysis*. Duxbury Press: Pacific Grove, CA.

Steven K. Thompson. 2002. *Sampling Statistics*. Chichester: New York.

▣ 원고 접수일 : 2005년 3월 17일
원고 심사일 : 2005년 6월 15일
심사 완료일 : 2005년 6월 15일